

Docket No. DE9-2000-0040 (269)

**METHOD AND APPARATUS FOR PHONETIC CONTEXT ADAPTATION FOR
IMPROVED SPEECH RECOGNITION**

Inventors:

Volker Fischer

Eric-W. Janke

Siegfried Kunzmann

A. Jon Tyrrell

International Business Machines Corporation

IBM Docket No. DE9-2000-0040

IBM Disclosure No. DE8-2000-0024

EXPRESS MAILING LABEL NO. EK972214875US

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of European Application No. 00124795.6, filed November 14, 2000 at the European Patent Office.

BACKGROUND OF THE INVENTION

1.1 Technical Field

The present invention relates to speech recognition systems, and more particularly, to a computerized method and apparatus for automatically generating from a first speech recognizer a second speech recognizer which can be adapted to a specific domain.

1.2 Description of the Related Art

To achieve necessary acoustic resolution for different speakers, domains, or other circumstances, today's general purpose large vocabulary continuous speech recognizers have to be adapted to these different situations. To do so, the speech recognizer must determine a huge number of different parameters, each of which can control the behavior of the speech recognizer. For instance, Hidden Markov Model (HMM) based speech recognizers usually employ several thousands of HMM states and several tens of thousands of multidimensional elementary probability density functions (PDFS) to capture the many variations of naturally spoken human speech. Therefore, the training of a highly accurate speech recognizer requires the reliable estimation of several millions of parameters. This is not only a time-consuming process, but also requires a substantial amount of training data.

It is well known that the recognition accuracy of a speech recognizer decreases significantly if the phonetic contexts and - in consequence of the changing phonetic contexts - pronunciations observed in the training data do not properly match those of the intended application. This is especially true when dealing with dialects or non-native speakers, but also can be observed when switching to other different domains,

EXPRESS MAILING LABEL NO. EK972214875US

for example within the same language or to other dialects. Commercially available speech recognition products try to solve this problem by requiring each individual end user to enroll in the system. Accordingly, the speech recognizer can perform a speaker-dependent re-estimation of acoustic model parameters.

5 Large vocabulary continuous speech recognizers capture the many variations of speech sounds by modelling context dependent sub-word units, such as phones or triphones, as elementary HMMs. Statistical parameters of such models are usually estimated from several hundred hours of labelled training data. While this allows a high recognition accuracy if the training data sufficiently represents the task domain, it can
10 be observed that recognition accuracy significantly decreases if phonetic contexts or acoustic model parameters are poorly estimated due to some mismatch between the training data and the intended application.

Since the collection of a large amount of training data and the subsequent training of a speech recognizer is both expensive and time consuming, the adaptation of a (general purpose) speech recognizer to a specific domain is a promising method to
15 reduce development costs and time to market. Conventional adaptation methods, however, either simply provide a modification of the acoustic model parameters or - to a lesser extent - select a domain specific subset from the phonetic context inventory of the general recognizer.

20 Facing both the industry's growing interest in speech recognizers for specific domains including specialized application tasks, language dialects, telephony services, or the like, and the important role of speech as an input medium in pervasive computing, there is a definite need for improved adaptation technologies for generating new speech recognizers. The industry is searching for technologies supporting the
25 rapid development of new data files for speaker (in-)dependent, specialized speech recognizers having improved initial recognition accuracy, and which require reduced customization efforts whether for individual end users or industrial software vendors.

SUMMARY OF THE INVENTION

One object of the invention disclosed herein is to provide for fast and easy customization of speech recognizers to a given domain. It is a further objective to provide a technology for generating specialized speech recognizers requiring reduced computation resources, for instance in terms of computing time and memory footprints. The objectives of the invention are solved by the independent claims. Further advantageous arrangements and embodiments of the invention are set forth in the respective dependent claims.

The present invention relates to a computerized method and apparatus for automatically generating from a first speech recognizer a second speech recognizer which can be adapted to a specific domain. The first speech recognizer includes a first acoustic model with a first decision network and corresponding first phonetic contexts. The present invention suggests using the first acoustic model as a starting point for the adaptation process. A second acoustic model with a second decision network and corresponding second phonetic contexts for the second speech recognizer can be generated by re-estimating the first decision network and the corresponding first phonetic contexts based on domain-specific training data.

Advantageously, the decision network growing procedure preserves the phonetic context information of the first speech recognizer which was used as a starting point. In contrast to state of the art approaches, the present invention simultaneously allows for the creation of new phonetic contexts that need not be present in the original training material. Thus, rather than create a domain specific inventory from scratch according to the state of the art, which would require the collection of a huge amount of domain-specific training data, according to the present invention, the inventory of the general recognizer can be adapted to a new domain based on a small amount of adaptation data.

BRIEF DESCRIPTION OF THE DRAWINGS

There are shown in the drawings, embodiments which are presently preferred, it being understood, however, that the invention is not so limited to the precise arrangements and instrumentalities shown.

5 Figure 1 is a flow diagram illustrating an exemplary structure for generating a speech recognizer which is tailored to a specific domain.

5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
220

DETAILED DESCRIPTION OF THE INVENTION

In the drawings and specification there is set forth a preferred embodiment of the invention, and although specific terms are used, the description thus given uses terminology in a generic and descriptive sense only and not for purposes of limitation.

5 The present invention can be realized in hardware, software, or a combination of hardware and software. Any kind of computer system - or other apparatus adapted for carrying out the methods described herein - is suited. A typical combination of hardware and software can be a general purpose computer system with a computer program that, when being loaded and executed, controls the computer system such that it carries out the methods described herein. The present invention also can be embedded in a computer program product, which comprises all the features enabling the implementation of the methods described herein, and which - when loaded in a computer system - is able to carry out these methods.

10 Computer program in the present context means any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after either or both of the following: a) conversion to another language, code or notation; b) reproduction in a different material form.

15 The present invention is illustrated within the context of the "ViaVoice" speech recognition system which is manufactured by International Business Machines Corporation, of Armonk, New York. Of course, the present invention can be used by any other type of speech recognition system. Moreover, although the present specification references speech recognizers which incorporate Hidden Markov Model (HMM) technology, the present invention is not limited only to such speech recognizers. Accordingly, the invention can be used with speech recognizers utilizing other approaches and technologies as well.

20 4.1 Introduction

Conventional large vocabulary continuous speech recognizers employ HMMs to compute a word sequence w with maximum a posteriori probability from a speech signal f . An HMM is a stochastic automaton $A = (\pi, \mathbf{A}, \mathbf{B})$ that operates on a finite set of states $S = \{s_1, \dots, s_N\}$ and allows for the observation of an output each time t , $t = 1, 2,$

..., T , a state is occupied. The initial state vector

$$\pi = [\pi_i] = [P(s(1) = s_i)], \quad 1 \leq i \leq N, \quad (\text{eq. 1})$$

gives the probabilities that the HMM is in state s_i at time $t=1$, and the transition matrix

$$\mathbf{A} = [a_{ij}] = [P(s(t+1) = s_j | s(t) = s_i)], \quad 1 \leq i, j \leq N, \quad (\text{eq. 2})$$

holds the probabilities of a first order time invariant process that describes the transitions from state s_i to s_j . The observations are continuous valued feature vectors

$\mathbf{x} \in \mathbb{R}$ derived from the incoming speech signal f , and the output probabilities are defined by a set of probability density functions (PDFS)

$$\mathbf{B} = [b_i] = [p(\mathbf{x} | s(t) = s_i)], \quad 1 \leq i \leq N. \quad (\text{eq. 3})$$

For any given HMM state s_i , the unknown distribution $p(\mathbf{x} | s_i)$ of the feature vectors is

approximated by a mixture of – usually gaussian – elementary probability density functions (pdfs)

$$\begin{aligned} p(\mathbf{x} | s_i) &= \sum_{j \in M_i} (\omega_{ji} \cdot N(\mathbf{x} | \mu_{ji}, \Gamma_{ji})) \\ &= \sum_{j \in M_i} (\omega_{ji} \cdot |2\pi\Gamma_{ji}|^{-1/2} \cdot \exp(-(\mathbf{x} - \mu_{ji})^T \Gamma_{ji}^{-1} (\mathbf{x} - \mu_{ji})/2)); \end{aligned} \quad (\text{eq. 4})$$

where M_i is the set of Gaussians associated with state s_i . Furthermore, \mathbf{x} denotes the observed feature vector, ω_{ji} is the j -th mixture component weight for the i -th output distribution, and μ_{ji} and Γ_{ji} are the mean and covariance matrix of the j -th Gaussian in

state s_i .

Large vocabulary continuous speech recognizers employ acoustic sub-word units, such as phones or triphones, to ensure the reliable estimation of a large number of parameters and to allow a dynamic incorporation of new words into the recognizer's vocabulary by the concatenation of sub-word models. Since it is well known that speech sounds vary significantly with respect to different acoustic contexts, HMMs (or HMM states) usually represent context dependent acoustic sub-word units. Moreover, since both the training vocabulary (and thus the number and frequency of phonetic contexts) and the acoustic environment (e.g. background noise level, transmission channel characteristics, and speaker population) will differ significantly in each target application, it is the task of the further training procedure to provide a data driven identification of relevant contexts from the labeled training data.

In a bootstrap procedure for the training of a speech recognizer, according to the state of the art, a speaker independent, general purpose speech recognizer is used for the computation of an initial alignment between spoken words and the speech signal. In this process, each frame's feature vector is phonetically labeled and stored together with its phonetic context, which is defined by a fixed but arbitrary number of left and/or right neighboring phones. For example, the consideration of the left and right neighbor of a phone P_0 results in the widely used (crossword) triphone context (P_{-1}, P_0, P_{+1}) .

Subsequently, the identification of relevant acoustic contexts (i.e. phonetic contexts that produce significantly different acoustic feature vectors) is achieved through the construction of a binary decision network by means of an iterative split-and-merge procedure. The outcome of this bootstrap procedure is a domain independent general speech recognizer. For that purpose some sets $Q_i = \{P_1, \dots, P_j\}$ of language and/or domain specific phone questions are asked about the phones at positions $K_{-m}, \dots, K_{-1}, K_{+1}, K_{+m}$ in the phonetic context string. These questions are of the

form: "Is the phone in position K_j in the set Q_i ?", and split a decision network node n into two successors, one node n_L (L for left side) that holds all feature vectors that give rise to a positive answer to a question, and another node n_R (R for right side) that holds the set of feature vectors that cause a negative answer. At each node of the network, the best question is identified by the evaluation of a probabilistic function that measures the likelihood $P(n_L)$ and $P(n_R)$ of the sets of feature vectors that result from a tentative split.

In order to obtain a number of terminal nodes (or leaves) that allow a reliable parameter estimation, the split-and-merge procedure is controlled by a problem specific threshold θ_p , i.e. a node n is split in two successors n_L and n_R , if and only if the gain in likelihood from this split is larger than θ_p :

$$P(n) < P(n_L) + P(n_R) - \theta_p \quad (\text{eq. 5})$$

A similar criterion is applied to merge nodes that represent only a small number of feature vectors, and other problem specific thresholds, e.g. the minimum number of feature vectors associated with a node, are used to control the network size as well.

The process stops if a predefined number of leaves is created. All phonetic contexts associated with a leaf cannot be distinguished by the sequence of phone questions that has been asked during the construction of the network, and thus are members of the same equivalence class. Therefore, the corresponding feature vectors are considered to be homogeneous and are associated with a context dependent, single state, continuous density HMM, whose output probability is described by a gaussian mixture model (eq. 4). Initial estimates for the mixture components are obtained by clustering the feature vectors at each terminal node, and finally the forward-backward algorithm known in the state of the art is used to refine the mixture component parameters. It is important to note, that according to this state of the art

procedure the decision network initially includes a single node and a single equivalence class only (refer to an important deviation with respect to this feature according to the present invention discussed below), which then iteratively is refined into its final form (or in other words the bootstrapping process actually starts "without" a pre-existing decision network).

In the literature, the customization of a general speech recognizer to a particular domain is known as cross domain modeling. The state of the art in this field is described for instance by R. Singh and B. Raj and R.M. Stern, "Domain adduced state tying for cross-domain acoustic modelling", Proc. of the 6th Europ. Conf. on Speech Communication and Technology, Budapest (1999), and roughly can be divided into two different categories:

1. extrinsic modeling: Here, a recognizer is trained using additional data from a (third) domain with phonetic contexts that are close to the special domain under consideration; and,
2. intrinsic modeling: This approach requires a general purpose recognizer with a rich set of context dependent sub-word models. The adaptation data is used to identify those models that are relevant for a specific domain, which is usually achieved by employing a maximum likelihood criterion.

While in extrinsic modeling one can hope that a better coverage of the application domain results in an improved recognition accuracy, this approach is still time consuming and expensive, because it still requires the collection of a substantial amount of (third domain) training data. On the other hand, intrinsic modeling utilizes the fact that only a small amount of adaptation data is needed to verify the importance of a certain phonetic context. However, in contrast to the present invention, intrinsic cross domain modeling allows only a fall back to coarser phonetic contexts (as this approach consists of a selection of a subset of the decision network and its phonetic context only), and is not able to detect any new phonetic context that is relevant to a new domain but not present in the general recognizer's inventory. Moreover, the

5 approach is successful only if the particular domain to be addressed by intrinsic modelling is already covered (at least to a certain extent) by the acoustic model of the general speech recognizer; or in other words, the particular new domain has to be an extract (subset) of the domain to which the general speech recognizer is already adapted.

4.2 Solution

10 If, in the following, the specification refers to a speech recognizer adapted to a certain domain, the term "domain" is to be understood as a generic term if not otherwise specified. A domain might refer to a certain language, a multitude of languages, a dialect or a set of dialects, a certain task area or set of task areas for which a speech recognizer might be exploited. For example, a domain can relate to certain areas within the science of medicine, the specific task of recognizing numbers only, and the like.

15 The invention disclosed herein can utilize the already existing phonetic context inventory of a (general purpose) speech recognizer and some small amount of domain specific adaptation data for both the emphasis of dominant contexts and the creation of new phonetic contexts that are relevant for a given domain. This is achieved by using the speech recognizer's decision network and its corresponding phonetic contexts as a starting point and by re-estimating the decision network and phonetic contexts based on domain-specific training data.

20 As the extensive decision network and the rich acoustic contexts of the existing speech recognizer are used as a starting point, the architecture of the proposed invention achieves minimization of both the amount of speech data needed for the training of a special domain speech recognizer, as well as the individual end users customization efforts. By upfront generation and adaptation of phonetic contexts towards a particular domain, the invention facilitates the rapid development of data files for speech recognizers with improved recognition accuracy for special applications.

The proposed teaching is based upon an interpretation of the training procedure of a speech recognizer as a two stage process that comprises 1.) the determination of relevant acoustic contexts and 2.) the estimation of acoustic model parameters. Adaptation techniques known the within the state of the art, for example maximum a posteriori adaptation (MAP) or maximum likelihood linear regression (MLLR), are directed only to the speaker dependent re-estimation of the acoustic model parameters ($\omega_{ji}, \mu_{ji}, \Gamma_{ji}$) to achieve an improved recognition accuracy; that is, these approaches exclusively target the adaptation of the HMM parameters based on training data. Importantly, these approaches leave the phonetic contexts unchanged; that is, the decision network and the corresponding phonetic contexts are not modified by these technologies. In commercially available speech recognizers, these methods are usually applied after gathering some training data from an individual end user.

In a previous teaching of V. Fischer, Y. Gao, S. Kunzmann, M. A. Picheny, "Speech Recognizer for Specific Domains or Dialects", PCT patent application EP 99/02673, it has been shown that upfront adaptation of a general purpose base acoustic model using a limited amount of domain or dialect dependent training data yields a better initial recognition accuracy for a broad variety of end users. Moreover it has been demonstrated by V. Fischer, S. Kunzmann, C. Waast-Ricard, "Method and System for Generating Squeezed Acoustic Models for Specialized Speech Recognizer", European patent application EP 99116684.4, that the acoustic model size can be reduced significantly without a large degradation in recognition accuracy based on a small amount of domain specific adaptation data by selecting a subset of probability density functions (PDFS) being distinctive for the domain.

Orthogonally to these previous approaches, the present invention focuses on the re-estimation of phonetic contexts, or - in other words - the adaptation of the recognizer's sub-word inventory to a special domain. Whereas in any speaker adaptation algorithm, as well as in the above mentioned documents of V. Fischer *et al.*,

the phonetic contexts once estimated by the training procedure are fixed, the present invention utilizes a small amount of upfront training data for the domain specific insertion, deletion, or adaptation of phones in their respective context. Thus re-estimation of the phonetic contexts refers to a (complete) recalculation of the decision network and its corresponding phonetic contexts based on the general speech recognizer decision network. This is considerably different from just "selecting" a subset of the general speech recognizer decision network and phonetic contexts or simply "enhancing" the decision network by making a leaf node an interior node by attaching a new sub-tree with new leaf nodes and further phonetic contexts.

The following specification refers to Fig. 1. Fig. 1 is a diagram reflecting the overall structure of the proposed methodology of generating a speech recognizer being tailored to a specific domain and gives an overview of the basic principle of the present invention. Accordingly, the description in the remainder of this section refers to the use of a decision network for the detection and representation of phonetic contexts and should be understood as but an illustration of one implementation of the present invention. The invention suggests starting from a first speech recognizer (1) (in most cases a speaker-independent, general purpose speech recognizer) and a small, i.e. limited, amount of adaptation (training) data (2) to generate a second speech recognizer (6) (adapted based on the training data (2)).

The training data (which is not required to be exhaustive of the specific domain) may be gathered either supervised or unsupervised, through the use of an arbitrary speech recognizer that is not necessarily the same as speech recognizer (1). After feature extraction, the data is aligned against the transcription to obtain a phonetic label for each frame. Importantly, while a standard training procedure according to the state of the art as described above starts the computation of significant phonetic contexts from a single equivalence class that holds all data (a decision network with one node only), the present invention proposes an upfront step that separates the additional data into the equivalence classes provided by the speaker independent, general purpose

speech recognizer. That is, the decision network and its corresponding phonetic contexts of the first speech recognizer are used as a starting point to generate a second decision network and its corresponding second phonetic contexts for a second speech recognizer by re-estimating the first decision network and corresponding first phonetic contexts based on domain-specific training data.

Therefore, for that purpose, the phonetic contexts of the existing decision network are first extracted as shown in step (31). The feature vectors and their associated phone context can be passed through the original decision network (3) by asking the phone questions that are stored with each node of the network to extract and to classify (32) the training data's phonetic contexts. As a result, one obtains a partitioning of the adaptation data that already utilizes the phonetic context information of the much larger and more general training corpus of the base system.

Subsequently, the original split-and-merge algorithm for the detection of relevant new domain specific phonetic contexts (4) can be applied resulting in a new, re-estimated (domain specific) decision network and corresponding phonetic contexts. Phone questions and splitting thresholds (refer for instance to eq. 5) may depend on the domain and/or the amount of adaptation data, and thus differ from the thresholds used during the training of the baseline recognizer. Similar to the method described in the introductory section 4.1, the procedure uses a maximum likelihood criterion to evaluate all possible splits of a node and stops if the thresholds do not allow a further creation of domain dependent nodes. This way one is able to derive a new, recalculated set of equivalence classes that can be considered by construction as a domain or dialect dependent refinement of the original phonetic contexts, which further may include, for HMMs associated with the leaf nodes of the re-estimated decision network, a re-adjustment of the HMM parameters (5).

One important benefit from this approach lies in the fact that - as opposed to using the domain specific adaptation data in the original, state of the art (refer for instance to section 4.1 above) decision network growing procedure - the present

invention preserves the phonetic context information of the (general purpose) speech recognizer which is used as a starting point. Importantly, and in contrast to cross domain modeling techniques as described by R. Singh *et al.* (refer to the discussion above), the method of the present invention simultaneously allows the creation of new phonetic contexts that need not be present in the original training material. Rather than create a domain specific HMM inventory from scratch according to the state of the art, which requires the collection of a huge amount of domain-specific training data, the present invention allows the adaptation of the general recognizer's HMM inventory to a new domain based on a small amount of adaptation data.

As the general speech recognizer's "elaborate" decision network with its rich, well-balanced equivalence classes and its context information is exploited as a starting point, the limited, i.e. small, amount of adaptation (training) data suffices to generate the adapted speech recognizer. This saves a significant effort in collecting domain-specific training data. Moreover, a significant speed-up in the adaptation process and an important improvement in the recognition quality of the generated adapted speech recognizer is achieved.

As with the baseline recognizer, each terminal node of the adapted (i.e. generated) decision network defines a context dependent, single state Hidden Markov Model for the specialized speech recognizer. The computation of an initial estimate for the state output probabilities (refer to eq. 4) has to consider both the history of the context adaptation process and the acoustic feature vectors associated with each terminal node of the adapted networks:

A. Phonetic contexts that are unchanged by the adaptation process are modelled by the corresponding gaussian mixture components of the base recognizer.

B. Output probabilities for newly created context dependent HMMs can be modelled either by applying the above-mentioned adaptation methods to the Gaussians of the original recognizer, or - if a sufficient number of feature vectors has been passed to the new terminal node - by clustering of the adaptation data.

Following the above mentioned teaching of V. Fischer *et al.*, "Method and System for Generating Squeezed Acoustic Models for Specialized Speech Recognizer", European patent application EP 99116684.4, the adaptation data may also be used for a pruning of Gaussians in order to reduce memory footprints and CPU time. The teaching of this reference with respect to selecting a subset of HMM states of the general purpose speech recognizer for use as a starting point ("Squeezing") and the teaching with respect to selecting a subset of probability-density-functions (PDFS) of the general purpose speech recognizer for use as a starting point ("Pruning"), both of which are distinctive of the specific domain, are incorporated herein by reference.

There are three additional important aspects of the present invention:

1. The application of the present invention is not limited to the upfront adaptation of domain or dialect-specific speech recognizers. Without any modification, the invention is also applicable in a speaker adaptation scenario where it can augment the speaker dependent re-estimation of model parameters. Unsupervised speaker adaptation, which requires a substantial amount of speaker dependent data, is an especially promising application scenario.

2. The present invention further is not limited to the adaptation of phonetic contexts to a particular domain (taking place once), but may be used iteratively to enhance the general recognizer's phonetic contexts incrementally based upon further training data.

3. If different languages share a common phonetic alphabet, the method also can be used for the incremental and data driven incorporation of a new language into a true multilingual speech recognizer that shares HMMs between languages.

4.3 Application Examples of the Present Invention

Facing the growing market of speech enabled devices that have to fulfill only a limited (application) task, the invention disclosed herein provides an improved recognition accuracy for a wide variety of applications. A first experiment focused on

the adaptation of a fairly general speech recognizer for a digit dialing task, which is an important application in the strongly expanding mobile phone market.

The following table reflects the relative word error rates for the baseline system (left), the digit domain specific recognizer (middle), and the domain adapted recognizer (right) for a general dictation and a digit recognition task:

| | baseline | digits | adapted |
|-----------|----------|--------|---------|
| dictation | 100 | 193.25 | 117.89 |
| digits | 100 | 24.87 | 47.21 |

The baseline system (baseline, refer to the table above) was trained with 20,000 sentences gathered from different German newspapers and office correspondence letters, and uttered by approximately 200 German speakers. Thus, the recognizer uses phonetic contexts from a mixture of different domains, which is the usual method to achieve good phonetic coverage in the training of general purpose, large vocabulary continuous speech recognizers, such as IBM's ViaVoice. The domain specific digit data included approximately 10,000 training utterances that further included up to 12 spoken digits and was used for both the adaptation of the general recognizer (adapted, refer to the table above) according to the teaching of the present invention and the training of a digit specific recognizer (digit, refer to the table above).

The above table gives the (relative) word error rates (normalized to the baseline system) for the baseline system, the adapted phone context recognizer, and the digit specific system. While the baseline system shows the best performance for the general large vocabulary dictation task, it yields the worst results for the digit task. In contrast, the digit specific recognizer performs best on the digit task, but shows unacceptable error rates for the general dictation task. The rightmost column demonstrates the benefits of the context adaptation: while the error rate for the digit recognition task

decreases by more than 50 percent, the adapted recognizer still shows a fairly good performance on the general dictation task.

4.4 Further Advantages of the Present Invention

5 The results presented in the previous section demonstrate that the invention described herein offers further significant advantages in addition to those addressed already within the above specification. From the discussion of the above outlined example, with respect to a general speech recognizer adapted to specific domain of a digit recognition task, it has been demonstrated that the present teaching is able to
10 significantly improve the recognition rate within a given target domain.

15 It has to be pointed out (as also made apparent by the above mentioned example) that the present invention at the same time avoids an unacceptable decrease of recognition accuracy in the original recognizer's domain. As the present invention uses the existing decision network and acoustic contexts of a first speech recognizer as a starting point, very little additional domain specific or dialect data, which is inexpensive and easy to collect, suffices to generate a second speech recognizer. Also
20 due to this chosen starting point, the proposed adaptation techniques are capable of reducing the time for the training of the recognizer significantly.

25 Finally, the invention allows the generation of specialized speech recognizers requiring reduced computation resources, for instance in terms of computing time and memory footprints. Accordingly, the invention disclosed herein is thus suited for the incremental and low cost integration of new application domains into any speech recognition application. It may be applied to general purpose, speaker independent speech recognizers as well as to further adaptation of speaker dependent speech recognizers. Still, the invention disclosed herein can be embodied in other specific forms without departing from the spirit or essential attributes thereof. Accordingly, reference should be made to the following claims, rather than to the foregoing specification, as indicating the scope of the invention.